Examiners' decision-making processes in observation-based clinical examinations

Malau-Aduli, Bunmi S.; Hays, Richard; D'Souza, Karen; Smith, Amy M.; Jones, Karina; Turner, Richard ; Shires, Lizzi; Smith, Jane W; Saad, Shannon; Richmond, Casandra; Celenza, Antonio; Sen Gupta, Tarun

Published in:
Medical Education

DOI:
10.1111/medu.14357

Licence:
Other

DR. BUNMI SHERIFAT MALAU-ADULI (Orcid ID : 0000-0001-6054-8498)

DR. KARINA JONES (Orcid ID : 0000-0001-7715-0507)

DR. SHANNON LEA SAAD (Orcid ID : 0000-0002-0423-5478)

Article type : Original Research

Examiners' decision-making processes in observation-based clinical examinations

Bunmi S. Malau-Aduli[1]*, Richard Hays[1], Karen D'Souza[2], Amy M. Smith[1], Karina Jones[1], Richard Turner[3], Lizzi Shires[3], Jane Smith[4], Shannon Saad[5], Cassandra Richmond[5], Antonio Celenza[6], Tarun Sen Gupta[1]

[1]College of Medicine and Dentistry, James Cook University, Townsville, Australia

[2]School of Medicine, Deakin University, Geelong, Australia

[3]School of Medicine, University of Tasmania, Hobart, Australia

[4]Medical Program, Bond University, Gold Coast, Australia

[5]School of Medicine, Notre Dame University, Sydney, Australia

[6]School of Medicine, University of Western Australia, Perth, Australia

*Corresponding author: Bunmi Malau-Aduli, E-mail: bunmi.malauaduli@jcu.edu.au

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/MEDU.14357

Abstract

Background:

Objective Structured Clinical Examinations (OSCE) are commonly used to assess the clinical skills of health professional students. Examiner judgement is one acknowledged source of variation in candidate marks. This paper reports an exploration of examiner decision-making to better characterise the cognitive processes and workload associated with making judgements of clinical performance in exit-level OSCEs.

Methods:

Fifty-five examiners for exit-level OSCEs at five Australian medical schools completed a NASA Task Load Index (TLX) measure of cognitive load and participated in focus group interviews immediately after the OSCE session. Discussions focused on how decisions were made for borderline and clear pass candidates. Interviews were transcribed, coded and thematically analysed. NASA TLX results were quantitatively analysed.

Results:

Examiners self-reported higher cognitive workload levels when assessing a borderline candidate in comparison to a clear pass candidate. Further analysis revealed five major themes considered by examiners when marking candidate performance in an OSCE: (a) Use of marking criteria as a source of reassurance; (b) Difficulty adhering to the marking sheet under certain conditions; (c) Demeanour of candidates; (d) Patient safety and (e) Calibration using a mental construct of the "*mythical* [prototypical] intern". Examiners demonstrated particularly higher mental demand when assessing borderline compared to clear pass candidates.

Conclusions:

Examiners demonstrate that judging candidate performance is a complex, cognitively difficult task, particularly when performance is of borderline or lower standard. At program exit level, examiners intuitively want to rate candidates against a construct of a prototypical graduate when marking criteria appear not to describe both what and how a passing candidate should demonstrate in completing clinical tasks. This construct should be shared, agreed upon and aligned with marking criteria to best guide examiner training and calibration. Achieving this integration may improve the accuracy and consistency of examiner judgements and reduce cognitive workload.

Introduction

Ensuring public trust in the competence of doctors is of paramount importance in medical education. Student progress decisions are usually based on results from several different assessment methods that are applied systematically to assess all desirable attributes of clinical knowledge and performance.[1] Each method should contribute to overall reliability, validity and impact on learning, withstanding scrutiny from regulators, employers and the public.[2] A common assessment method is the Objective Structured Clinical Examination (OSCE), where examiners mark candidates performing a sample of clinical tasks in a series of standardized encounters or "stations".[3-5] Examiners mark candidate performance against a checklist and/or global rating scale and, ideally, provide written feedback. OSCE results often contribute significantly to progress decisions, making more important the reliability of examiner judgements.[3,6]

Variations in assessment scores should reflect differences in candidate performances, but several potential sources of error exist, including station, patient and examiner characteristics.[7-11] Minimising judgment error is particularly important due to the relatively small number of stations.[12-15] Examiner judgements have received substantial recent attention.[7-10,16] Documented OSCE examiner variance is up to four times that of examinee variance[8,12,17,18], with potentially substantial impact on pass/fail rates.[8,19,20] Discrepancies in the global ratings and pass/fail decisions have been reported between examiners.[7,21] Differences persist despite examiner training and improved station design[22-24], highlighting a need to understand how examiners make judgements.[25-28]

Assessing clinical performance is a complex cognitive process that involves initial impressions, active detection and selection of relevant performance elements, and then processing, assimilation and categorisation in working memory.[29] Retrieval of integrated information from long-term memory informs understanding of the required standard.[29,30] An attribute that is easier to evaluate may be used when judging performance instead of an intended attribute.[27,31-33] Examiners may be less confident to make a fail decision than a pass, and may alter a judgement when given limited additional information that contradicts more robust information.[34] Additionally, information processing theory suggests that other factors, such as the number of performance elements exhibited by the candidate, how the information was presented (i.e. order, organisation, completeness, quality and speed), and associated processing requirements (e.g. patient complexity) make judgement more difficult.[35] These findings are particularly important when judging borderline candidates, where there may be a fine line between 'minimally competent' and 'just failing' performances, yet substantial consequences.[36]

Three perspectives on examiner cognition from educational psychology are helpful in understanding this complexity: heuristics, natural decision making, and social cognition theory. Heuristics are unconscious cognitive processes that may reduce the strain associated with managing multiple cognitively demanding tasks.[37,38] Natural decision making involves the use of schemas which are automatic, based on stable, prior, long-term knowledge.[39-43] Both approaches may improve efficiency, but are rapid and use fewer cues and so are prone to bias.[32,37,44] In social cognition theory, decisions are influenced by the motivations of individual decision makers and the local practices wherein the decision making takes place.[45]. All three perspectives may help understand how to improve assessment quality.[27]

Sadler (1989)[46] described two major cognitive theories about complex decision making: analytic and configurational. Analytic approaches combine information to reach an overall measure, whereas configurational approaches form an initial holistic impression and then substantiate it with reference to multiple criteria, often 'fuzzy' rather than 'sharp'. A competent judge identifies relevant criteria, but not necessarily in advance.[46] Recognition-primed decisions are more frequent than analytical decisions, particularly with experienced judges, in time-pressured situations or when judgements are difficult.[39]

Previous exploratory studies on examiner cognition have focused on workplace based assessments (WBA).[40,47-51] Given the prominence of the OSCE format in medical education, it is important to explore examiner cognition in the context of time-limited, multi-station assessments.

The differing cognitive processes of examiners in formulating judgements raises questions about the impact of cognitive workload on making complex judgements.[52,53] Increased cognitive load may decrease activity in particular regions of the brain needed to make social judgments[54], suggesting that workload might vary between clear pass and borderline judgements. Studying cognition during complex tasks is difficult; assumptive methodological techniques such as the NASA-TLX have been used to measure mental workload[55], although may require triangulation with other methods to increase understanding of relevance to medical education.[55-58]

There is scope for further research into the cognition of examiners in order to improve examiner training. In this paper we report an exploration of examiner thinking when judging examinee performance in high-stakes OSCEs. This information may assist both in refining examiner training – sharpening the examiner – and improvement in rating instrument design – sharpening the tool.

Methods

Study Context

We focused on exit level OSCEs of Australian medical schools given the requirement for common curriculum/assessment mapping to agreed national graduate outcomes, determined by the Australian Medical Council.[59] Participating schools are members of the Australian Collaboration for Clinical Assessment in Medicine (ACCLAiM) and have similar integrated, outcomes-based curricula and similar OSCE processes, including some shared stations, examiner calibration exercises and similar scoring sheets consisting of a checklist and a global rating scale, with explicit performance category descriptors.[60,61] This study was approved by the James Cook University Human Research Ethics Committee (H6833) and accepted by all participating universities.

Study Design

A pragmatic, mixed-methods approach explored the cognitive processes examiners follow when making judgements about candidate performance. The different strengths of qualitative (focus group) and quantitative (NASA TLX cognitive load index) approaches were applied, adopting the core principle that experiences cannot be separated from the social contexts in which they occur[62], to provide an integrated picture of examiner cognition when making complex judgments. Data collection was immediately after OSCE sessions, as close as possible to candidate interactions.

Participant recruitment

Nine of the twelve ACCLAiM member schools were holding exit-level OSCEs at the time of this study and were invited to participate by an email sent at least four weeks prior to the OSCE. Examiners at each school were invited to participate via email one week in advance, and again at the examiner briefing session. Volunteering participants were provided an information sheet and signed an informed consent form. Consent was confirmed verbally at each group discussion.

Cognitive load

The NASA Task Load Index (NASA TLX) measures subjective mental workload across six dimensions: mental demand; physical demand; temporal demand; performance; effort; and frustration.[63] Participants place an X on a 21-point line for each dimension. At commencement of the focus group, participants were instructed: “Please cast your mind back to a candidate you assessed in the last session who was a clearly passing candidate. Please now complete the NASA TLX scale based on assessing that candidate.” When

that task was completed, the same instructions were given, this time reflecting on a borderline candidate. This approach was intended to facilitate richer responses from participants during the focus group phase.

Focus group sessions

Focus groups were conducted because of convenience (maximising recruitment in a short window after the OSCE) and the capacity to enhance clarification of participants' views by exploring how and why they think in a particular way.[62] Three of the authors (BMA, KDS and RBH), all external to the host school, moderated the group discussions, using the same agreed interview schedule with probing questions based on their experience as examiners and informed by the literature (see Appendix 1).

All sessions commenced with information about the purpose of the study, and reflection on a clearly passing and a borderline candidate. The sessions proceeded with questions about examiners' experiences marking the OSCEs, focusing on the two candidate performances, using the same set of questions for both scenarios. Participants were able to react to and build upon each other's responses to build a deeper understanding of the issues discussed. Responses were clarified and expanded upon with follow-up probing questions. The moderator ensured that all contributed to the discussions. Sessions lasted between 45-60 minutes.

Sampling and Data recording

The numbers of focus groups and participants were based on a purposive sampling framework, but there were elements of convenience, as not all schools could participate within the time frame and only examiners able to remain after the examinations could participate. Participants were de-identified and differentiated by gender and a participant ID. All focus groups were recorded and transcribed verbatim by professional transcribers. Data collection and analysis occurred concurrently and ceased after five schools as focus group responses were no longer revealing new information.[64]

Quantitative data analysis

Data from the NASA TLX were entered into IBM SPSS Statistics Version 25 software to determine whether differences in self-reported mental workload were statistically significant when assessing "clear pass" versus "borderline" candidates on all six dimensions. X placements on the 21-point line were interpreted as being 10-point numeric scale ranging from 0 (no mental demand/perfect) to 10 (very high demand/fail); data were recorded at 0.25 increments and raw TLX responses were analysed. Data were not normally distributed, so the non-parametric Wilcoxon Signed Rank Test was used, with statistical significance deemed at the alpha = 0.05 level. The association between a NASA TLX dimension and assessment outcomes was assessed using the effect size calculation for Wilcoxon Signed Rank Tests[65], and

comparing this with Cohen's (1988)[66] guidelines for determining small (0.1), medium (0.3), or large (0.5) effects. This provided a quantitative description of the six dimensions; differences were used to further qualify focus group findings.

Qualitative data analysis

Focus group transcripts were analysed thematically using NVivo Plus Version 12 (QSR International Pty Ltd., 2018). A cyclical, iterative and reflective process was applied to coding the data to identify themes, compare similarities and differences in responses and to describe the processes examiners used to make OSCE assessment judgements. The coding process was completed by two of the researchers (AMS and BMA) in four phases.[64] In the initial phase, transcripts were read and coded line-by-line, for specific mentions of (1) OSCE performance characteristics that influenced judgments, (2) general strategies examiners used to make judgments, and (3) challenges examiners experienced in their roles as examiners. In the next 3 phases, codes were grouped into themes, coding definition was reviewed and emerging themes were compared and integrated with the pre-existing literature.[64] Subsequently, emerging themes were confirmed by two other authors (KDS and RBH) and discrepancies were resolved in a consensus meeting.

Results

Sample

A total of 55 examiners, 27 male and 28 female, participated in seven focus groups at five schools (Table 1). Their median (IQR) years of OSCE examining experience was 10 (5-15) years. Their 'usual' roles and varied relationships with the candidates, from regular lecturers who knew the candidates well to part-time clinicians with varying experience levels and no knowledge of individual candidates.

Table 1. Demographic characteristics of focus group participants.

Cognitive Load

Forty-seven examiners (85% of participants) rated their workload using the NASA TLX. Higher cognitive workload levels ($\alpha = 0.5$) on all six workload dimensions were reported for assessing a borderline candidate compared to a clear pass candidate (Table 2). The magnitude of these effects ranged from medium (0.3) to large (0.5).

Table 2. Descriptive statistics and results of Paired Samples significance tests (Clear Pass - Borderline)

The qualitative analysis was consistent with the cognitive workload results, with examiners reporting increased mental effort when marking borderline candidates. For many, feeling a sense of struggle was a sign they were marking a borderline candidate.

> "*I always struggle with borderline…. and having to do the agonising" – FG2, Female*

> *"But those ones who are in the middle, are much harder because they start off, and they are not that good and you have to concentrate really well, because you want to know if they are just that side or that side, and…their problem is…sometimes they get most of the stuff, but their structure is not very good, so you can't just go, tick, tick, tick, tick. It's like, they're, they're all over the place, and you have to make a decision, are they…just that side of the line or just that side of the line, and sometimes I think that's quite hard" – FG7, Male*

Examiners felt the need to record justification when giving borderline or fail marks. Many reported writing detailed notes and candidate feedback. Time constraints made this challenging, as both making borderline judgments and providing written justification take time. One examiner explained a need to finish comments for one student while another student was performing; this detracted from the attention on the current student.

> *"I find the borderline ones really difficult, 'cos I'm afraid that, you know, what the impact would be on them of giving them a borderline, so, I find it much more taxing to make the final decision. And it's not always, it's not clear, there is often a mixed results, in all the little things that you tick off, but it's an overall impression which is sometimes an emotional decision, you know, and so, that's what worries me. Because we are all supposed to be trying to be, you know, um, non-emotional and objective, but sometimes, it's just a feeling that you've got"* - FG6, Female

Thematic analysis

Thematic analysis of the focus group transcripts revealed five major themes in relation to how examiners make judgements about OSCE candidate performance: (a) Use of marking criteria as reassurance; (b) Difficulty adhering to the marking sheet under certain conditions; (c) Demeanour of candidates; (d) Patient safety and (e) Calibration using a mental construct of a 'prototypical intern'.

Marking criteria as reassurance - the 'safety blanket'

Examiners mainly adhered to station-specific marking criteria, because this assisted with standardisation and reassurance that their judgements were supported. This was especially true for borderline candidates, where an examiner reported that justification was likely to be sought.

> *"I was constantly ticking off on the hard paper thing when they said certain questions or investigations, so that was a really great safety blanket" – FG3, Female.*

Increasing examiner consistency was a strong motivation for adhering to the marking criteria. A calibration meeting was regarded as important for building this confidence by helping examiners interpret and mark the form consistently.

> *"I think it's much easier to stick with the marking sheets and abide by them if you have a very good pre-meeting before going in" – FG1, Female*

> *"I pretty much stuck very much with the process, largely because of confidence in the process. So with enough years of confidence to say, well, if I stick with it [laughs], I should get a reasonable consistency" – FG3, Male.*

Difficulty adhering to the marking sheet under certain conditions

OSCE marking criteria were seen as helpful for identifying presence/absence of required attributes, but the criteria sometimes lacked important assessable nuances. Where desired marking criteria were absent or under a different heading, personal judgement was used.

> *"I'm ticking you according to the sheet, but there's just something that's not [accounted for on the sheet]. [To address that] I would usually, because I have to go by the criteria ……probably score them down a little bit and make a comment…because I don't feel that they should get the same score as someone who, just given that really clear, um, sort of understanding" – FG7, Female.*

Consistency was especially challenging when station content did not seem to align with the marking criteria.

> *"Additional questions or prompts were sometimes needed to ensure students understood what was being asked of them. Modifications are also needed to be made when, for example, a patient did not want to proceed in the scripted way. [It's a] challenge when you need to deviate. If we go off script, then that's not comparable with the other stations. So we really can't go off* script" – FG5, Male.

Demeanour of candidates

Candidate demeanour was important, including the perceived level of engagement, decisiveness, poise and confidence in performing the task. Behaviours and attitudes that demonstrated empathy, fluency and self-awareness were viewed favourably by examiners and used to differentiate higher from lower performing candidates, although were not explicit in marking criteria.

> *"To me, that part of my marking really comes down to how confident I feel that that student will - you know, if that student was my resident, well, how confident would I be in what they were doing and would I need to be constantly checking?" – FG2, Female.*

Patient Safety

Examiners frequently considered patient safety, particularly for borderline candidates, asking themselves:

> *"Will the patients be safe based on this student's behaviour?"* – FG7, Female

Examiners placed such high importance on safety that an assessment of 'not safe' can be the determining factor in a 'fail' outcome for the station. Examiners reported it easier to make a 'fail' judgement based on safety considerations than for other deviations from expectations.

> *"I think, sometimes with a fail…I've got that thing in my mind about, you know, community responsibility, and if they are…really not up to the mark, the best thing for them and the community is not to pass them. And, you know, if they are not safe, then… I think that's somewhat reassuring" - FG7, Female*

Considerations of safety were important enough to motivate examiners to deviate from the marking sheet.

> *"The problem is that there are some inherent parts there about safety, et cetera, that are really pertinent to the case, which, unfortunately, wasn't actually recognised well in the rubric. That being said… you do need to actually ask yourself… can this person with the actual knowledge they demonstrate in this actual station here, would they be actually a safe person to look after community?" – FG2, Male*

> *"You know, if that student was my resident, well, how confident would I be in what they were doing and would I need to be constantly checking. So it comes down - or how confident would I be if they were looking after my family member" - – FG1, Female.*

Calibration using the 'prototypical' intern

Sometimes examiners employed other strategies in which they compared candidate performances against their own 'norm-referenced' standards. When setting these standards, some examiners reported using the concept of the *"mythical [prototypical] intern*" (introduced by a FG2, Male 1), which was their perception of what the typical intern should be, based on a personally-derived set of standards that was used as comparison.

> *"The first person you are marking is actually compared with that mythical [prototypical] intern"* – FG2, Male.

> *"I think I construct a mythical [prototypical] intern. But I often think if I was the registrar on the phone to that intern, or coming into this situation, you know, I sort of think about how that would be playing out. So I sort of put myself in the role play a little bit, I think. I guess there's sort of memories of what it was like to be an intern and working with other interns. So there's sort of a little bit of a personal paradigm attached to the mythical [prototypical] intern, if you like"* – FG2, Female.

The standard was sometimes based on an examiner's evaluations of what they would do as a doctor, intern, and/or at their stage of education.

> "*I guess, basically, I compare them to myself. What would I do in practice?"* (FG2, Male)

Some examiners, particularly those with strong clinical backgrounds, based their standard on how they expected junior staff to perform.

> *"I probably didn't do it on personal experience, So I just thought about how I want my juniors to talk to their staff, and that did come into play as well, like that degree of precepts that I want them to be able to convey"* (FG4, Female)

There were cases of examiners trying to remember what it was like to be in the examinees' shoes.

> *"But I must say that the failing is not a bad thing because that will give them more training and focus and they actually probably do better than the one who passed, because I failed my clinical exam …... They give me a lot of chances to interact with a lot of people and think about things a bit broader and deeper, and maybe I feel a better physician"* (FG2, Male)

As the OSCE continued, the point of comparison moved from personal constructs to direct comparison with earlier candidates. When asked why they started with a pre-formed concept and then changed to comparison with early candidates, examiners gave reasons of needing "*time to think about it because I*

don't want to rush to judgement" (FG6, Male), *"you need a sample"* (FG6, Male), and "*after a while you get an idea"* (FG5, Male).

What versus how

Generally, the examination and borderline/pass judgment processes followed a similar general strategy for all respondents. Examiners initially followed the marking criteria, looking for presence/absence of clearly identified assessment items to measure *what* the candidate did. However, the global score required an assessment of not only *what,* but also *how*.

> *The students are capable of ticking the rubric box and still demonstrating that they're disorganised or right in their thought processes. So at the end of the day, they may well have ticked enough boxes to pass but it's our judgement that they are borderline. Therefore, there is a box at the end that is our overall overview, and that's the one that we really need to soul search on. Probably the OSCE process doesn't look at that well enough* – FG2, Male.

When making judgments of *how*, examiners discussed a variety of strategies used to differentiate a passing candidate from a borderline or failing candidate. These strategies incorporated both analytical (criterion-based) and configurational (gut feeling/affective) approaches. Analytical techniques included: making conscious comparisons with other reference sources (including training aids and marking criteria). Affective considerations related to the level of confidence/comfort the examiner had with how this candidate could perform on their own, often based on the examiner's personal concept of the prototypical intern. A recognition that no single individual examiner can pass or fail a candidate overall eased the emotional turmoil caused by the prospects of making a borderline or fail judgment for their station. Examiners were aware of an association between high performance and better 'acting' and the ability to perform well under scrutiny. They tried to make allowances but there was scope for personal values and weighting of these performance aspects to influence judgments.

> *"I think the second thing was that he timed it so well that he knew what was relevant and what was important for the purpose of the exam and the content matter"* – FG1, Female.

Discussion

This study provides rich detail regarding the pragmatic and affective considerations that examiners employ when assessing observed clinical performance at an OSCE station. The finding that examiner decision making is a substantial cognitive task is consistent with results from other studies.[40,47-51,67] This study delves deeper into the cognitive strategies that examiners use when judging clinical performance in OSCEs, providing both qualitative and quantitative data. The cognitive workload is higher when

candidates are marked as a borderline or fail performance, with examiners describing feelings as a "*struggle" and "agonising*". These self-perceptions were flagged by examiners as subjective evidence that the observed performance was likely to be borderline. The reported decreased confidence and increased mental effort when marking borderline performance correlates well with the results of the NASA TLX that confirmed high mental demand and frustration with marking the borderline compared to the clear pass performance. Examiners often used all the available time at the station to settle upon a final mark of borderline candidate. Such candidates may exhibit "patchy" performance – competent in some tasks, but not in others – making it difficult to reach an overall judgement. Examiners were concerned about the defensibility of lower ratings, so tried to make detailed comments to justify their decisions, even though time constraints made this difficult. Occasionally, marking is completed during the next station iteration, potentially reducing concentration on the next candidate's performance. This may have implications for station timing and the number of stations or duration of examining assigned to each examiner.

Educational psychology theories seem to fit quite well with how OSCE examiners make judgements. They try to be analytical by following checklists and criteria, relying on these criteria as a *"safety blanket"* to check that content was either present or absent, particularly for borderline candidates.[39] Examiners were aware of the need to ensure consistency amongst all examiners marking the same station – *"sticking with the process"*. While the checklist provided the *"what"*, the global judgement scale was used as a holistic assessment of *"how"* tasks were performed, involving both practical *"head-based"* and *"instinctive"* considerations. Examiners sometimes asked unscripted additional questions, deviated from the marking criteria or used the global score to correct for perceived poor alignment between the expected tasks and the marking criteria "*there is a box at the end that is our overall overview, and that is the one we really need to soul search on".* While experienced examiners generally make more accurate global ratings[4,22,68], some examiners in this study were relatively inexperienced. The potential influence of contrasting with prior students and the resultant biasing effects may lead to variations in the use of and confidence in global rating scales and this could ultimately affect standard setting.

Examiners often fall back on heuristics and schema for the first candidate in the session and for more difficult decisions when marking criteria do not align well with their expectations of sound performance.[37,42,43] This explains the concept of the prototypical intern, based on personal experiences and expectations particularly to judge *how* skills were performed. Because we did not investigate decisions about individual candidates, we could not explore the impact of social cognition, which may be relevant when there is such variability in workplace relationships between examiners and candidates. Appropriate student demeanour is regarded by examiners as characteristic of a high-performing student,

highlighting the potential for a 'halo effect'.[69-72] A candidate may score well by exhibiting confidence and efficiency.[73] For clearly passing candidates, this may be due to stronger clinical experience, but could a weak candidate appear experienced and yet not have sufficient knowledge?

A feature of the prototypical intern is the focus on readiness and safety for practice as a junior doctor. Candidates were compared against this personal, internal, implied standard, almost despite the marking criteria. If candidate performance suggested a need for constant checking ("how confident would I be if they were looking after my family member?"), then marks were lower. This resonates with work reported by Crossley and Jolly[74], supporting readiness for practice as a clinician-aligned construct that is not always explicit in marking criteria.

Strengths/ limitations of study

Strength of this study include the multi-institutional, nationwide participation and the close temporal relationship between examining and data collection. Limitations include the possibility that only confident examiners took part and that recall bias affected focus group discussions. Memory interference from ensuing candidates may have occurred due to the gap between remembered candidates. Future areas for study include the impact of aligning marking criteria with examiners' cognitive frameworks and standardisation of the prototypical intern in examiner calibration.

Conclusion/ Take home message

This study provides an exploration of what is happening 'inside the black box' of OSCE examiners' minds when rating students' performances in exit-level, presenting concordance between the examiner's narrative with objective measures of cognitive load. The process of making judgements about candidate performance is complex and cognitively difficult, combining analytical and affective elements. Particularly when exit-level candidate performance is borderline or lower, examiners intuitively want to rate candidates against a personal construct of a prototypical graduate when marking criteria appear not to describe both the 'what' and the 'how' of candidate performance. This construct should be shared, agreed and aligned with marking criteria to guide examiner training and calibration. Achieving this integration may improve the accuracy and consistency of examiner judgements and may also reduce cognitive workload and/or increase efficiency of examiner decision-making.

Contributors: BMA, RBH, KDS and TSG conceived the study. BMA, RBH and KDS conducted the focus groups. BMA, AMS, KJ, RH and KDS advised on data analysis and interpretation and contributed to writing

the original draft of the manuscript. All authors facilitated collection of data. All authors edited, reviewed and accepted the final version of the manuscript.

Acknowledgements: The authors gratefully acknowledge the contributions of the examiners who participated in this study. We also thank the administrative and academic staff from all participating schools who facilitated the organisation of the focus groups.

Funding: No external funding required

Declaration of interest: The authors report no conflicts of interest.

Ethical approval: Overarching ethics approval was obtained from JCU. In addition, all participating schools obtained ethics approval from their local Ethics Committee. All information was de-identified before data analysis.

Appendix 1: Focus group questions

1. What it was about that performance that allowed you to make the judgement?

2. With what were you comparing the performance?

3a. Were there some aspects that were not so good but were compensated for by excellence elsewhere? **(for pass candidate only)**

3b. Were there some omissions or errors that were compensated for by substantial effort elsewhere? **(for borderline candidate only)**

4. How confident were you in deciding upon the rating for this examinee?

5. How much time was needed to reach a judgement/efficiency?

6. How stable was your initial judgement? Did you change your ratings?

7. To what degree did you adhere to the procedures/forms?

8. What level of influence has your experience in teaching (medical students, trainees, IMGs) had on your examiner role today?

References

1. Norcini J, Anderson MB, Bollela V, et al. 2018 Consensus framework for good assessment. *Medical teacher.* 2018;40(11):1102-1109.
2. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med.* 2008;40(8):574-578.
3. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education.* 1979;13(1):39-54.
4. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Medical Education.* 2003;37(11):1012-1016.
5. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education.* 2004;38(2):199-203.
6. Downing SM, Yudkowsky R. *Assessment in health professions education.* New York: Routledge; 2009.
7. Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE. A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Academic Medicine.* 2005;80(10):S59-S62.
8. Harasym PH, Harasym PH, Woloschuk W, Woloschuk W, Cunning L, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advances in Health Sciences Education.* 2008;13(5):617-632.
9. Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. *Medical Education.* 2009;43(6):526-532.
10. Bartman I, Smee S, Roy M. A method for identifying extreme OSCE examiners. *The Clinical Teacher.* 2013;10(1):27-31.
11. Yeates P, Moreau M, Eva K. Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects? *Academic Medicine.* 2015;90(7):975-980.
12. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality Assurance Methods for Performance-Based Assessments. *Advances in Health Sciences Education.* 2003;8(1):27-47.
13. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education.* 2004;38(3):327-333.
14. Messick S. Validity. In: Linn RL, ed. *Educational Measurement.* 3rd ed. New York: American Council on Education, Macmillan Publishing; 1889:13-103.

15. Downing SM. Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education.* 2005;39(4):353-355.
16. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Medical Education.* 2011;45(12):1181-1189.
17. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education.* 2006;6(1):42-42.
18. Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Medical Education.* 2010;44(7):690-698.
19. Brennan PAMDF, Croke DTP, Reed MMDF, et al. Does Changing Examiner Stations During UK Postgraduate Surgery Objective Structured Clinical Examinations Influence Examination Reliability and Candidates' Scores? *Journal of Surgical Education.* 2016;73(4):616-623.
20. Hope D, Cameron H. Examiners are most lenient at the start of a two-day OSCE. *Medical Teacher.* 2015;37(1):81-85.
21. Wong ML, Fones CSL, Aw M, et al. Should non-expert clinician examiners be used in objective structured assessment of communication skills among final year medical undergraduates? *Medical Teacher.* 2007;29(9-10):927-932.
22. Malau-Aduli BS, Mulcahy S, Warnecke E, et al. Inter-rater reliability: comparison of checklist and global scoring for OSCEs. 2012.
23. Reid K, Smallwood D, Collins M, Sutherland R, Dodds A. Taking OSCE examiner training on the road: reaching the masses. *Med Educ Online.* 2016;21(1):32389-32385.
24. Schleicher I, Leitner K, Juenger J, et al. Examiner effect on the objective structured clinical exam - A study at five medical schools. *BMC Medical Education.* 2017;17(1):71-77.
25. Chong L, Taylor S, Haywood M, Adelstein B-A, Shulruf B. The sights and insights of examiners in objective structured clinical examinations. *Journal of educational evaluation for health professions.* 2017;14:34.
26. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. *Medical Education.* 2016;50(5):511-522.
27. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Medical Education.* 2014;48(11):1055-1068.
28. Govaerts MJB, Govaerts MJB, van der Vleuten CPM, et al. Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Advances in Health Sciences Education.* 2007;12(2):239-260.

29. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education.* 2013;18(2):291-303.
30. Yaphe J, Street S. How do examiners decide?: A qualitative study of the process of decision making in the oral examination component of the MRCGP examination. *Medical Education.* 2003;37(9):764-771.
31. Chahine S, Holmes B, Kowalewski Z. In the minds of OSCE examiners: uncovering hidden assumptions. *Advances in Health Sciences Education.* 2016;21(3):609-625.
32. Gilovich T, Griffin DW. Heuristics and biases: then and now. In: Gilovich T, Kahneman D, Griffin DW, eds. *Heuristics and biases: the psychology of intuitive judgment.* Cambridge: Cambridge University Press; 2002.
33. Klein G. The recognition-primed decision (RPD) model: Looking back, looking forward. *Naturalistic decision making.* 1997:285-292.
34. Tweed MJ, Thompson-Fawcett M, Wilkinson TJ. Decision-making bias in assessment: The effect of aggregating objective information and anecdote. *Medical Teacher.* 2013;35(10):832-837.
35. Baddeley AD. *Working memory, thought, and action.* Oxford: Oxford University Press; 2007.
36. Burr SA, Zahra D, Cookson J, Salih VM, Gabe-Thomas E, Robinson IM. Angoff anchor statements: setting a flawed gold standard? *MedEdPublish.* 2017;6(3).
37. Tversky A, Tversky A, Kahneman D, Kahneman D. Judgment under uncertainty: heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty. *Science.* 1974;185(4157):1124-1131.
38. Connolly T, Arkes HR, Hammond KR. *Judgment and decision making: An interdisciplinary reader, 2nd ed.* New York, NY, US: Cambridge University Press; 2000.
39. Klein G. Naturalistic Decision Making. *Human Factors: The Journal of the Human Factors and Ergonomics Society.* 2008;50(3):456-460.
40. Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. *Advances in health sciences education : theory and practice.* 2012;18(4):559-571.
41. Govaerts MJ, Schuwirth LW, Vleuten CPMvd, Muijtjens AMM. Workplace-based assessment: effects of rater expertise. *Advances in health sciences education : theory and practice.* 2011;16(2):151-165.
42. Govaerts MJ, Wiel MWJvd, Schuwirth LW, Vleuten CPMvd, Muijtjens AMM. Workplace-based assessment: raters' performance theories and constructs. *Advances in health sciences education : theory and practice.* 2013;18(3):375-396.

43. Shah AK, Oppenheimer DM. Heuristics Made Easy: An Effort-Reduction Framework. *Psychological Bulletin.* 2008;134(2):207-222.
44. Gigerenzer G, Gaissmaier W. Heuristic Decision Making. *Annual review of psychology.* 2011;62(1):451-482.
45. Levy PE, Williams JR. The Social Context of Performance Appraisal: A Review and Framework for the Future. *Journal of Management.* 2004;30(6):881-905.
46. Sadler DR. Formative assessment and the design of instructional systems. *Instructional science.* 1989;18(2):119-144.
47. Gingerich A, Vleuten CPMvd, Eva KW, Regehr G. More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic medicine.* 2014;89(11):1510-1519.
48. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model: Opening the black box of direct observation. *Medical education.* 2011;45(10):1048-1060.
49. Roduta Roberts M, Cook M, Chao ICI. Exploring assessor cognition as a source of score variability in a performance assessment of practice-based competencies. *BMC medical education.* 2020;20(1):168-114.
50. Yeates P, Yeates P, O'Neill P, et al. Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education.* 2013;18(3):325-341.
51. Lee V, Brain K, Martin J. From opening the 'black box' to looking behind the curtain: cognition and context in assessor-based judgements. *Advances in Health Sciences Education.* 2019;24(1):85-102.
52. Bodenhausen GV, Morales JR. Social cognition and perception. *Handbook of Psychology, Second Edition.* 2012;5.
53. Tavares W, Eva KW. Impact of rating demands on rater-based assessments of clinical competence. *Education for Primary Care.* 2014;25(6):308-318.
54. Jenkins AC. Rethinking Cognitive Load: A Default-Mode Network Perspective. *Trends in Cognitive Sciences.* 2019;23(7):531-533.
55. Paravattil B, Wilby KJ. Optimizing assessors' mental workload in rater-based assessment: a critical narrative review. *Perspect Med Educ.* 2019;8(6):339-345.
56. Charles RL, Nixon J. Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics.* 2019;74:221-232.

57. Gingerich A, Yeates P. The mental workload of conducting research in assessor cognition. *Perspect Med Educ.* 2019;8(6):315-316.
58. Naismith LM, Cavalcanti RB. Validity of Cognitive Load Measures in Simulation-Based Training: A Systematic Review. *Academic Medicine.* 2015;90(11 Association of American Medical Colleges Medical Education Meeting: Proceedings of the 54th Annual Research in Medical Education Sessions):S24-S35.
59. Australian Medical Council. Accreditation and Recognition. https://www.amc.org.au/accreditation-and-recognition/assessment-accreditation-primary-medical-programs/. Published 2018. Accessed 03 May 2019.
60. Malau-Aduli BS, Teague P-A, D'Souza K, et al. A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools. *Medical Teacher.* 2017;39(12):1261-1267.
61. Malau-Aduli BS, Teague P-A, Turner R, et al. Improving assessment practice through cross-institutional collaboration: An exercise on the use of OSCEs. *Medical Teacher.* 2016;38(3):263-271.
62. O'Leary Z. *The essential guide to doing your research project.* 3rd ed. London, United Kingdom;Thousand Oaks, California;: SAGE Publications; 2017.
63. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: *Advances in psychology.* Vol 52. Elsevier; 1988:139-183.
64. Birks M, Mills J. *Grounded theory: a practical guide.* Second ed. Thousand Oaks, CA;London;: SAGE; 2015.
65. Pallant JF. *SPSS survival manual: a step by step guide to data analysis using IBM SPSS.* 6th ed. Sydney, New South Wales: Allen & Unwin; 2016.
66. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, N.J: L. Erlbaum Associates; 1988.
67. Yeates P, Cardell J, Byrne G, Eva KW. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. *Medical Education.* 2015;49(9):909-919.
68. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Academic Medicine.* 1999;74(10):1129-1134.
69. Iramaneerat C, Iramaneerat C, Yudkowsky R, et al. Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education.* 2008;13(4):479-493.
70. Iramaneerat C, Yudkowsky R. Rater Errors in a Clinical Skills Assessment of Medical Students. *Evaluation & the Health Professions.* 2007;30(3):266-283.

71. Wood TJ, Wood TJ. Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education.* 2014;19(3):409-427.
72. Wood TJ, Wood TJ, Chan J, et al. The influence of first impressions on subsequent ratings within an OSCE station. *Advances in Health Sciences Education.* 2017;22(4):969-983.
73. Chan M, Bax N, Woodley C, Jennings M, Nicolson R, Chan P. The first OSCE; does students' experience of performing in public affect their results? *BMC medical education.* 2015;15(1):59.
74. Crossley J, Jolly B. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical education.* 2012;46(1):28-37.

Table 1. Demographic characteristics of focus group participants.

School	Focus Group	*N*	Males (*n*)	Females (*n*)
S1	FG1	7	2	5
S2	FG2	13	8	5
S3	FG3	6	2	4
S3	FG4	2	0	2
S4	FG5	5	1	4
S5	FG6	12	8	4
S5	FG7	10	6	4
	Total	55	27	28

Table 2. Descriptive statistics and results of Paired Samples significance tests (Clear Pass - Borderline)

Dimension	Outcome	Descriptive Statistics			Wilcoxon Signed Rank Test[a]		
		n	Median	IQR[b]	*z*	*p*	*r*[c]
Mental Demand	Clear Pass	47	5.50	4.75	-5.41	<.01	0.56
	Borderline	47	7.75	2.00			
Physical Demand	Clear Pass	46	1.75	2.31	-2.39	0.017	0.25
	Borderline	45	2.00	3.75			
Temporal Demand	Clear Pass	47	5.00	3.75	-3.33	<.01	0.34
	Borderline	47	6.75	2.75			
Performance	Clear Pass	47	2.50	2.00	-3.74	<.01	0.39
	Borderline	46	3.50	2.50			
Effort	Clear Pass	47	5.00	4.50	-4.27	<.01	0.44
	Borderline	47	7.25	2.75			
Frustration	Clear Pass	47	1.75	1.50	-5.55	<.01	0.57
	Borderline	47	5.00	3.25			

[a] Based on negative ranks

[b] IQR = interquartile range

[c] *r* = effect size